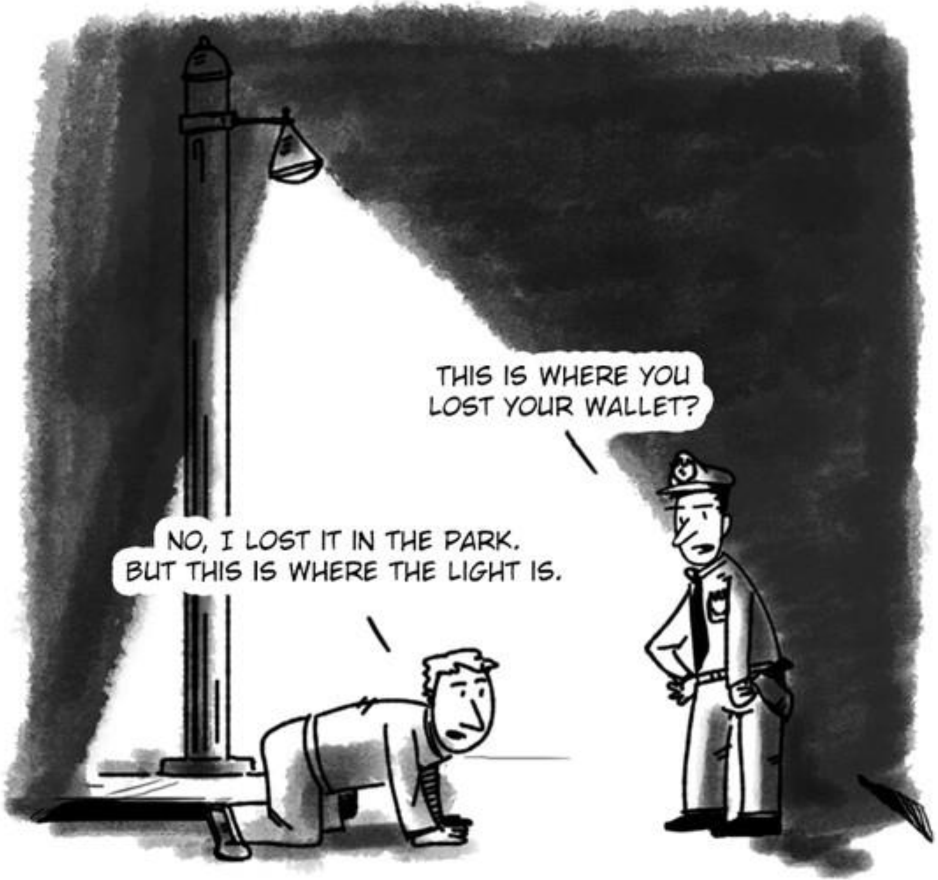
Analyzing large-scale genomics data

Bas Heijmans
Molecular Epidemiology
Leiden University Medical Center
The Netherlands
bas.heijmans@lumc.nl

FOS course Molecular Data Science – 22 November 2021.

The drunkard's search effect


THIS IS WHERE YOU LOST YOUR WALLET?
NO, I LOST IT IN THE PARK. BUT THIS IS WHERE THE LIGHT IS.

From 1 to all

- All genetic variants, genes, metabolites
  → comprehensive & representative
  (instead of generalizing a single bit of knowledge)
- Disease ≠ 1 gene
  → hypotheses (!) and discoveries on the full complexity of biology.
- Exploiting natural variation
  → The human as model organism

Learning objectives

1. SPSS 2nd
2. ‘*R*’ 1st

Why?

From traditional data to large-scale (high-dimensional) data:

- Many different formats of data files
- Data require preprocessing (quality control, normalization)
- Many tests (thousands, millions, billions)
- Novel methods
- Computationally intensive methods
- Smart figures to make sense of data
- Visualizations to make sense of results
- Linking to external knowledge for interpretation

How to in SPSS

GeneExpression.cel
(Affymetrix)

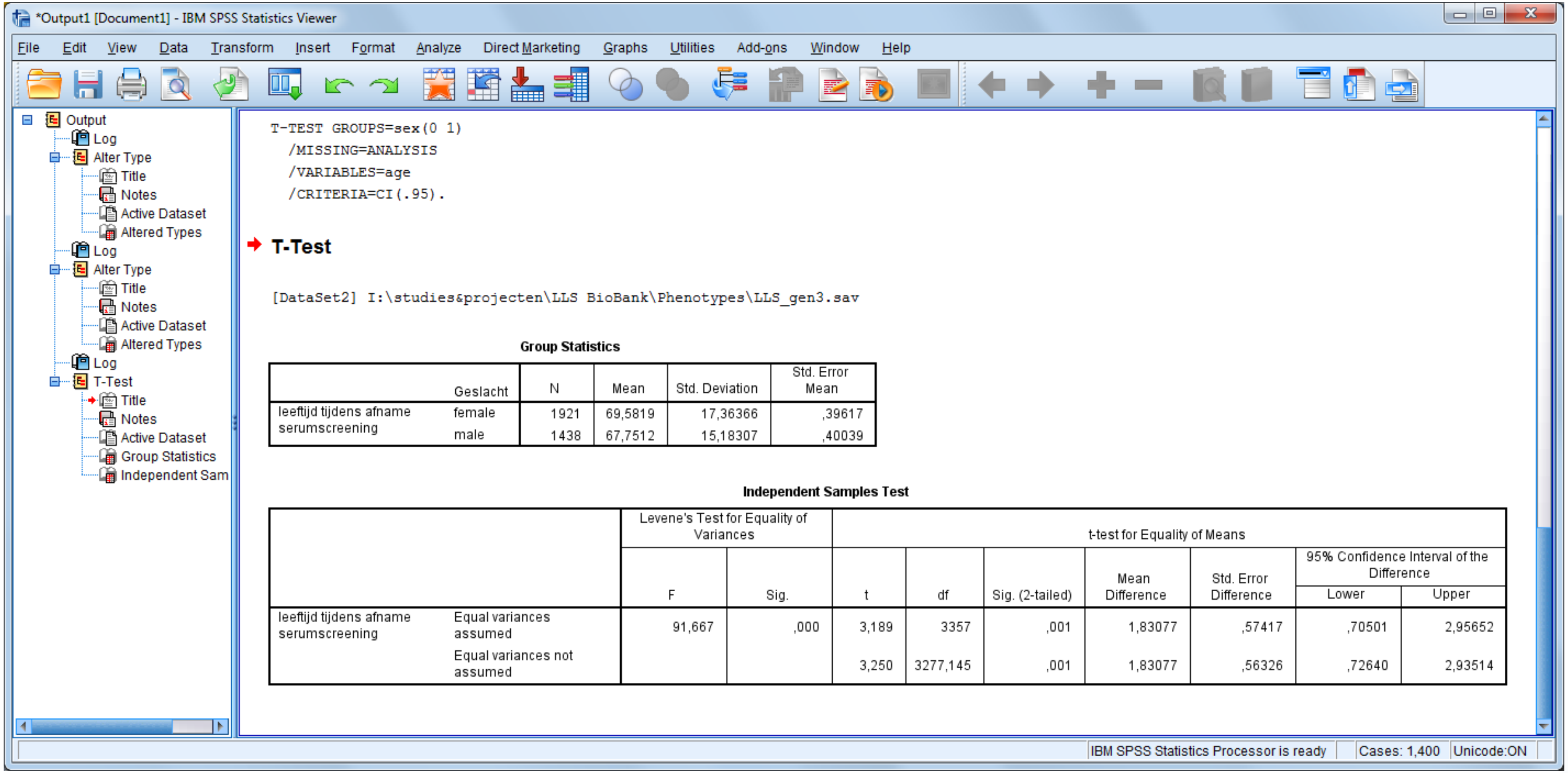
GeneExpression.idat
(Illumina)

- Many different formats of data files.
- Data require preprocessing (quality control, normalization) prior to analysis.

How to in SPSS

Person-id	Expression gene 1	Outcome
1	10	2.3
2	6	0.9
…	…	….
1000	15	1.5

How to in SPSS



```
T-TEST GROUPS=sex(0 1)
  /MISSING=ANALYSIS
  /VARIABLES=age
  /CRITERIA=CI(.95).
```

T-Test

```
[DataSet2] I:\studies&projecten\LLS BioBank\Phenotypes\LLS_gen3.sav
```

Group Statistics

	Geslacht	N	Mean	Std. Deviation	Std. Error Mean
leeftijd tijdens afname serumscreening	female	1921	69,5819	17,36366	,39617
	male	1438	67,7512	15,18307	,40039

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
leeftijd tijdens afname serumscreening	Equal variances assumed	91,667	,000	3,189	3357	,001	1,83077	,57417	,70501	2,95652
	Equal variances not assumed			3,250	3277,145	,001	1,83077	,56326	,72640	2,93514

How to in SPSS

Person-id	Gene 1	Gene 2	…	Gene 22,703	Outcome
1	10	1		90	2.3
2	6	0		54	0.9
…	…	…		…	….
1000	15	3		39	1.5

- 22,703 tests
- Repeat same analysis many times and store results in one data object.

How to in SPSS

Person-id	Variant 1	Variant 2	…	Variant $7x10^6$	Gene 1	Gene 2	…	Gene 22,703	Outcome
1	AG	TT		AT	10	1		90	2.3
2	GG	TC		AA	6	0		54	0.9
…	…	…	…	…	…	…	…	…	…
1000	GG	TC		AT	15	3		39	1.5

- 7M x 22,703 tests
- Distribute computations across processors (parallelization)
- Novel methods

How to in SPSS

Person-id	Variant 1	Variant 2	…	Variant 7x10^6	Gene 1	Gene 2	…	Gene 22,703	Outcome
1	AG	TT		AT	10	1		90	2.3
2	GG	TC		AA	6	0		54	0.9
…	…	…	…	…	…	…	…	…	…
1000	GG	TC		AT	15	3		39	1.5

- 7M x 22,703 tests
- Smart figures to make sense of data
- Visualizations to make sense of results
  → 0.1 trillion (= 10^{11}) p-values

How to in SPSS

Person-id	Variant 1	Variant 2	…	Variant $7x10^6$	Gene 1	Gene 2	…	Gene 22,703	Outcome
1	AG	TT		AT	10	1		90	2.3
2	GG	TC		AA	6	0		54	0.9
…	…	…	…	…	…	…	…	…	…
1000	GG	TC		AT	15	3		39	1.5

- 7M x 22,703 tests
- Linking to external knowledge for interpretation (e.g. location variant, function of gene)

How to in SPSS

Person-id	Expression gene 1	Outcome
1	10	2.3
2	6	0.9
…	…	….
1000	15	1.5

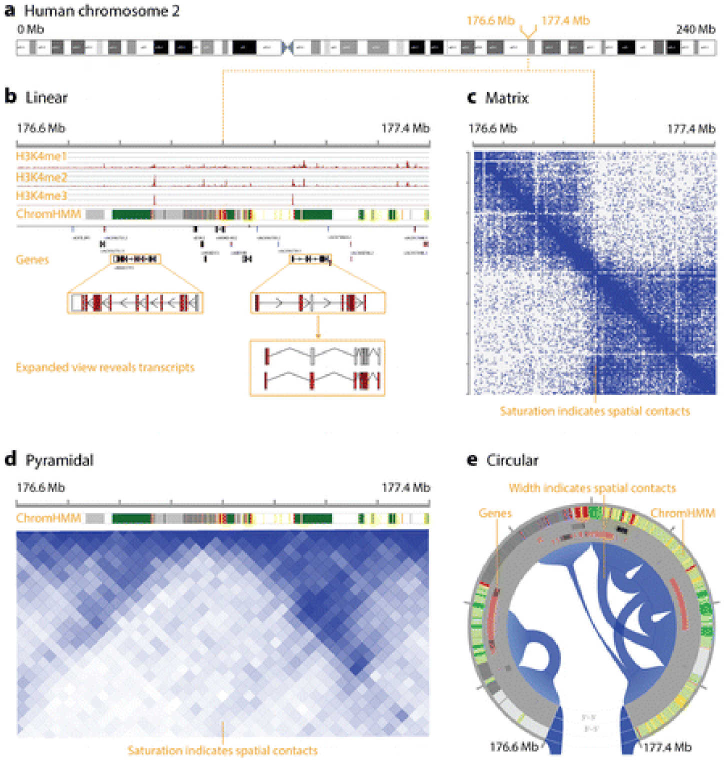
- Click-fest
- Complex output
- Ugly graphs
- Black-box: need to trust developers

Why?

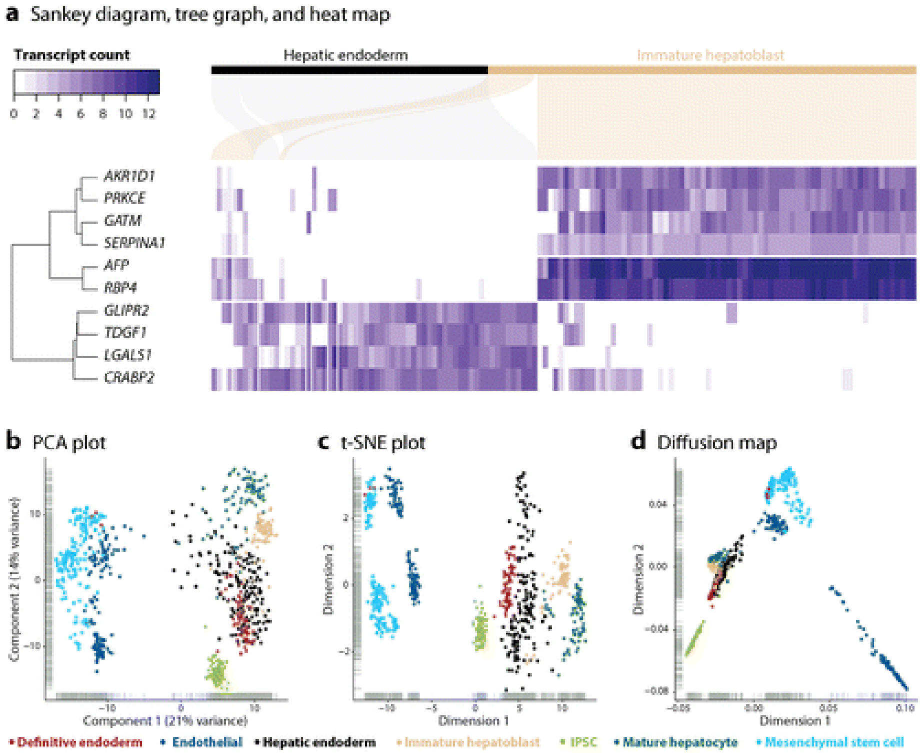
From traditional data to large-scale (high-dimensional) data

- Many different formats of data files
- Data require preprocessing (quality control, normalization)
- Many tests (thousands, millions, billions)
- Novel methods
- Computationally intensive methods
- Smart figures to make sense of data
- Visualizations to make sense of results
- Linking to external knowledge for interpretation

Visualizations



O'Donoghue SI, et al. 2018.
Annu. Rev. Biomed. Data Sci. 1:275–304



O'Donoghue SI, et al. 2018.
Annu. Rev. Biomed. Data Sci. 1:275–304

O'Donoghue et al. Annu Rev Biomed Data Sci 2018

R first

- Do not fear the blinking cursor!
- You will find that R is not more complicated than SPSS if scripts are available.
- But: some analyses you will do are!
- Curriculum in transition: this is not an R course (a flavour of R & not all is in R).
- Also: R is not the answer to all issues in bioinformatics.